VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INTELLIGENT SYSTEMS

ANALÝZA CHOVÁNÍ UŽIVATELŮ V PROSTŘEDÍ BEZDRÁTOVÝCH SÍTÍ

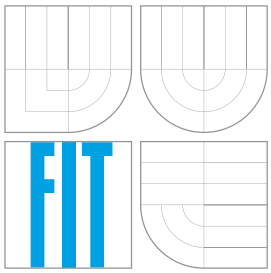
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

MICHAL JACKO

BRNO 2014

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF INTELLIGENT SYSTEMS

ANALÝZA CHOVÁNÍ UŽIVATELŮ V PROSTŘEDÍ BEZDRÁTOVÝCH SÍTÍ

ANALYSIS OF USER BEHAVIOR IN THE WIRELESS NETWORKS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — MICHAL JACKO
AUTHOR

VEDOUCÍ PRÁCE — Ing. MATEJ KAČIC
SUPERVISOR

BRNO 2014

Abstrakt

Táto práca sa zaoberá problémom analýzy správania používateľov v prostredí bezdrôtových sietí. Práca popisuje návrh a implementáciu metódy, ktorá na základe správania dokáže klasifikovať jednotlivých používateľov.

Abstract

This paper deals with a problem of analysys of user behavior in the wireless networks. Paper describes design and implementation of a method, which can classify users by their behavior.

Klíčová slova

Wi-Fi, bezpečnost sítí, chování uživatelů, klasifikace, dolování z dat

Keywords

Wi-Fi, network security, user behavior, classification, data mining

Citace

Michal Jacko: Analýza chování uživatelů v prostředí bezdrátových sítí, bakalářská práce, Brno, FIT VUT v Brně, 2014

Analýza chování uživatelů v prostředí bezdrátových sítí

Prohlášení

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne pod vedením pána Ing. Mateja Kačica.

.....
Michal Jacko
31. července 2014

Poděkování

Týmto chcem poďakovať Ing. Matejovi Kačicovi za vedenie práce a cenné rady pri jej riešení. Taktiež by som chcel poďakovať Ing. Ivanovi Homoliakovi za odbornú pomoc a konzultácie, ktoré mi poskytol.

Obsah

Kapitola 1

Úvod

Bezdrôtové siete sa v dnešnej dobe stávajú bežnou možnosťou prístupu do počítačových sietí a siete Internet. Tieto siete sú omnoho náchylnejšie na útoky než siete využívajúce pevné prenosové médium, pretože útočník nepotrebuje fyzický prístup k sieťovej infraštruktúte. Väčšina bezdrôtových sietí je dnes zabezpečená pomocou štandardu 802.11i [4], ktorý je považovaný za bezpečný. Mnoho útokov, hlavne vo firemných sieťach, je však vykonávaných z vnútra siete, kedy je útočník legitímny užívateľ pripojený do siete.

Cieľom práce je vytvoriť nástroj, ktorý identifikuje každú entitu v bezdrôtovej sieti a na základe predošlého správania danej entity dokáže detekovať zmeny v správaní, ktoré môžu poukazovať na potencionány útok na sieť.

Práca je členená do niekoľkých kapitol. V nasledujúcej kapitole je rozobrané zabezpečenie bezdrôtových sietí a možnosti útokov na ne. Ďalšia kapitola obsahuje analýzu problému chovania používateľov v sietiach. V kapitole 4 je popísaný spôsob riešenia danej problematiky. Nasleduje kapitola popisujúca spôsob testovania a dosiahnuté výsledky. Táto kapitola taktiež obsahuje návrh možných rozšírení do budúcnosti. V záverečnej kapitole sa nachádza zhrnutie dosiahnutých výsledkov.

Kapitola 2

Bezdrôtové siete Wi-Fi

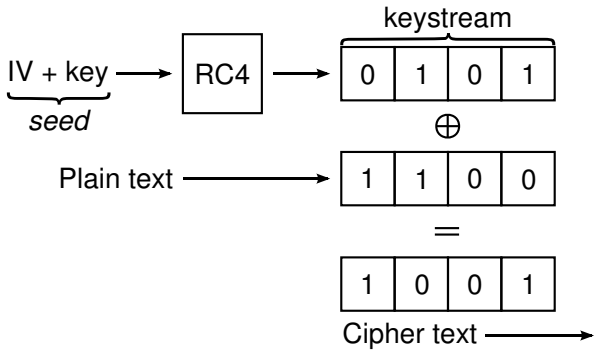
Wi-Fi je rozšírené a často používané označenie štandardu IEEE 802.11[4]. Je to protokol pre sieťovú komunikáciu, ktorý pracuje na prvej a druhej vrstve ISO/OSI modelu. Tento štandard bol prijatý roku 1997. Na jeho základe boli v nasledujúcich rokoch prijaté ďalšie štandardy z rodiny 802.11x, ktoré zdokonalovali, rozširovali a dopĺňali pôvodnú špecifikaciu. Prenos dát je uskutočňovaný elektromagnetickým vlnením na voľnom frekvenčnom pásme 2,4 GHz, prípadne v pásme 5 GHz. Protokol definuje najmä prenosové rýchlosti, frekvenčné pásma, formát dátových rámcov, fyzickú adresáciu sieťových prvkov a riešenie kolízii pri prenose. Keďže je prenos bezdrôtový, sila signálu je závislá na množstve a type prekážok medzi vysielačom a prijímačom. Čím viac prekážok, tým je väčšia strata signálu a následne aj zníženie rýchlosti prenosu. V praxi sú teda prenosove rýchlosti približne polovičné ako oficiálne uvádzané rýchlosti, ktoré môžu byť dosiahnuté iba za ideálnych podmienok, ako je priama viditeľnosť a žiadne kolízie na zdielanom médiu.

2.1 Zabezpečenie bezdrôtových sietí

Bezdrôtové siete sú vo všeobecnosti považované za náchylnejšie na útoky než siete využívajúce pevné prenosové médium, pretože útočník nepotrebuje fyzický prístup k sieťovej infraštruktúte. Rádiové siete možno ľahko odpočúvať, preto musia mať zabudované mechanizmy pre zaistenie prístupu iba pre oprávnených užívateľov(autentizácia) a taktiež mechanizmy pre zabezpečenie samotnej komunikácie (šifrovanie a zabezpečenie integrity prenášaných dát). V nasledujúcej kapitole boli informácie čerpané z [5].

2.1.1 Filtrácia MAC adries

Jedným z nie príliš účinných spôsobov zabezpečenia Wi-Fi sietí je filtrovanie MAC adries. Zabezpečenie týmto spôsobom spočíva v tom, že je vytvorený zoznam MAC adries, ktoré majú možnosť sa do siete pripojiť. Ak sa do siete pokúsi pripojiť zariadenie s adresou MAC, ktorá nie je na zozname, zariadeniu je prístup odmietnutý. Hodnota MAC adresy zariadenia je však ľahko modifikovateľná, tento spôsob zabezpečenia preto nepredstavuje pre útočníka veľkú prekážku. Odpočúvaním prevádzky útočník zistí, ktoré adresy sú v sieti povolené. Keď sa zariadenie s určitou MAC adresou od siete odpojí, zmení útočník svoju MAC adresu na adresu odpojeného zariadenia a prejde cez filter MAC adries. Filtrácia MAC adries sa preto používa ako sekundárne zabezpečenie k iným spôsobom zabezpečenia.



Obrázek 2.1: šifrovanie WEP

2.1.2 WEP

Schéma zabezpečenia WEP bola navrhnutá v pôvodnej špecifikácii 802.11 pre dosiahnutie takej bezpečnosti komunikácie v bezdrôtovýchch sietiach, aká odpovedá bezpečnosti u pevných sietí LAN. Schéma WEP využíva algoritmus RC4 na šifrovanie dát a CRC na zabezpečenie integrity. Šifrovanie dát môže byť realizované pomocou 64 alebo 128 bitového kľúča, ktorý vznikne zreťazením tajného klúča (40 alebo 104 bitov) a inicializačného vektora (24 bitov), ktorý sa mení obvykle s každým paketom. Schéma postupu šifrovania je znázornená na obrázku 2.1.

2.1.3 WPA

Zabezpečenie pomocou WEP bolo považované za nedostatočné, preto bolo potrebné zaviesť nový spôsob zabezpečenia Wi-Fi sietí. WPA (Wi-Fi Protected Access) je dočasné riešenie a predstavuje prechod medzi technológoiu WEP a 802.11i (WPA2). Pri návrhu WPA bola zohľadnená požiadavka, aby nebolo potrebné na zabezpečenie vykonávať zmenu v existujúcom HW.

WPA ponúka viacero možností autentizácie užívateľov. V prípade malých sietí je najviac využívaný spôsob autentizácie pomocou predzdieľaného kľúča (PSK). V tomto prípade zdieľajú všetky stanice heslo pozostávajúce z 8 až 63 ASCII znakov, prípadne 64 hexadecimálnych číslic. Ďalším spôsobom autentifikácie je mód Enterprise, pri ktorom je potrebná dostupnosť centralizovaného autentizačného servra, ktorý je väčšinou implementovaný ako RADIUS server. Autentizačný proces je popísaný pomocou protokolu EAP.

Ako šifrovací algoritmus je z dôvodu spätnej kompatibility použitý RC4. Používa sa šifrovací kľúč dĺžky 128 bitov pozostávajúci z inicializačného vektoru o dĺžke 48 bitov a 80-bitového tajného kľúča. Na rozdiel od protokolu WEP sa kľúč mení s každým paketom, čím je odolný proti útokom hrubou silou. Schéma takéhoto šifrovania sa nazýva TKIP (Temporal Key Integrity Protocol).

Na zaistenie integrity dát sa využíva kód MIC (Message Integrity Code). Ku každému rámcu je pridaný digitálny podpis, zabraňujúci útočníkovi odchytiť paket, zmeniť ho a poslať ďalej.

2.1.4 802.11i(WPA2)

Štandard 802.11i známy tiež ako WPA2 bol prijatý v roku 2004. Bol navrhnutý na poskytnutie čo najväčšieho stupňa zabezpečenia. Jeho nevýhodou je spätná nekompatibilita so

staršími zariadeniami, ktoré nemajú postačujúci HW pre pokročilé šifrovanie.

Šifrovanie aj zabezpečenie integrity je založené na šifrovacom algoritme AES (Advanced Encryption Standard), ktorý je v súčastnosti považovaný za vysoko bezpečný štandard [5]. Protokol WPA2 používa modifikáciu AES v podobe CCMP protokolu. Na rozdiel od WPA a WEP poskytuje zabezpečenie integrity celého rámcu vrátane hlavičiek. Podobne ako pri WPA, kľúče sú dynamicky menené pre každý prenášaný paket. Vo WPA2 sa nevyužíva inicializačný vektor, ale čísla paketov.

Autentizácia používateľov je v štandarde WPA2 rovnaká ako v prípade WPA.

2.2 Útoky na bezdrôtové siete

Bezdrôtové siete sú náchylné na rôzne druhy útokov, ktoré je možné rozdeliť na dve základné kategórie - pasívne a aktívne. K pasívnym útokom patrí odpočúvanie siete a analýza prevádzky. Medzi aktívne možno zaradiť opakovanie, modifikáciu správ, odmietnutie služby (DoS) a falšovanie identity.

2.2.1 DoS útoky

Denial of Service je technika útoku, pri ktorej dochádza k zahlteniu požiadavkami a nedostupnosti služby pre ostatných používateľov. Dôvodom pre daný typ útoku môže byť snaha útočníka o zresetovanie prístupového bodu, čo môže byť využité pri útokoch man-in-the-middle.

Útokov typu DoS existuje veľké množstvo, medzi tie najznámejšie patria ICMP flood, SYN flood, Deauth flood či RTS/CTS DoS.

2.2.2 Útok man-in-the-middle

Posdtatou tohto útoku je snaha útočníka odpočúvať komunikáciu medzi účastníkmi tak, že sa stane aktívnym prostredníkom. Sfalšovaním MAC adries oklame útočník prístupový bod, pre ktorý sa tvári ako autorizovaný klient a taktiež je oklamaný klient, ktorému útočník predstiera, že je autorizovaný prístupový bod. Táto pozícia môže byť náseldne zneužitá na odpočúvanie, prípadne modifikáciu komunikácie medzi oklamanými zariadeniami.

2.2.3 Útoky na WEP

Najjednoduchším, avšak nie príliš účinným útokom na WEP je útok hrubou silou zameraný na prelomenie tajného kľúča. Tento typ útoku nie je prakticky použiteľný pri dlhších heslách, kedy otestovanie všetkých možností trvá príliš dlhú dobu, v závislosti na výpočetnej sile útočníka. Tento útok môže byť kombinovaný so slovníkovou metódou, ktorou sa zmenší počet skúšaných kombinácii avšak nie je zaručené prelomenie hesla.

Efektívnejší a rozšírenejší typ útoku je FMS útok [5], pomenovaný podľa iniciálov jeho tvorcov (Fluhrer, Mantin, Shamir). Tento útok využíva zraniteľnosti algoritmu RC4 - slabé inicializačné vektory (IV, na základe ktorých je možné zistiť hodnotu tajného kľúča), generovanie IV s rovnakou hodnotou a predpokladané hodnoty prvých bytov nešifrovaných dát. Využitým týchto znalostí stačí útočníkovi zachytiť dostatočné množstvo paketov obsahujúcich slabé IV, z ktorých môže odvodiť tajný kľúč. Tento typ útoku je často sprevádzaný útokom typu DoS, vďaka ktorému je zvýšená prevádzka na sieti, čím je znížený čas potrebný na zachytenie dostatočného množstva paketov.

2.2.4 Útoky na WPA2

V súčastnosti nie sú náme žiadne efektívne a prakticky vykonateľné útoky, ktoré by využívali chyby v šifrovacom algoritme AES použitom vo WPA2.

V roku 2010 bola objavená bezpečnostná zraniteľnosť známa ako Hole 196. Nejde o prelomenie v zmysle odhalenia privátneho kľúča, teda prienik zvonku do šifrovanej WPA2 siete. Útok sa týka len prostredia Enterprise, útočník musí byť legitímne autetnifikovaný do šifrovanej WPA2 siete. Po EAP autentifikácii nasleduje tzv „4-way handshake“ medzi stanicou a AP, pri ktorom sa vygeneruje kľúč PTK (Pairwise Transient Key), ktorý slúži na šifrovanie unicastovej komunikácie (medzi stanicou a AP). Pri generovaní tohto kľúča je zahrnutá MAC adresa zariadenia, čo zabezpečuje ochranu voči jej podvrhnutiu. Túto vlastnosť však nemá GTK kľúč, ktorý sa používa na šifrovanie multicastovej a brodcastovej komunikácie. V praxi to znamená, že útočník môže vytvoriť vlastný broadcast rámec s podvrhnutou adresou AP a vydávať sa za pôvodné AP, pričom stanice to nepoznajú. Úspešné využitie tejto chyby umožní útočníkovi vykonať útok typu man-in-the-middle.

Kapitola 3

Analýza problému

Mnoho používateľov sa spolieha na bezpečnosť štandardu WPA2 a jeho šifrovací algoritmus AES. Veľké množstvo útokov však prebieha zvnútra sietí, kedy je útočníkom legitímny používateľ siete. Cieľom mojej práce je analýza chovania každého používateľa v sieti a na jej základe vytvorenie modelu správania každej entity v danom prostredí. Na základe tohto modelu je možné identifikovať zmenu správania danej entity, ktorá môže poukazovať na potencionálnu nežiadanú činnosť, či útok na sieť.

Chovanie používateľa v sieti je komplexný proces, avšak v činnosti užívateľa existujú mnohé pravidelnosti, ktoré je možné odhaliť. Každý človek má určité návyky, ktoré sa odrážajú v správaní v bezdrôtových sietiach. Správanie používateľov je možné klasifikovať na základe rôznych vlastností, napríklad:

- čas, v ktorom komunikuje na sieti
- prehliadanie webových stránok (využitie protokolu HTTP a HTTPS, počty navštívených stránok)
- e-mailová komunikácia (využitie rôznych protokolov (IMAP alebo POP3), šifrovanej/nešifrovanej komunikácie)
- práca s vzdialenými zariadeniami pomocou protokolu SSH
- multicastová komunikácia

Jednou z možností na analýzu správania užívateľov je sledovanie štatistík činnosti daného používateľa v sieti. Cieľom je vytvoriť pre každého užívateľa ku každému dňu graf, ktorý obsahuje hodnoty jednotlivých vlastností zachytených v jednotlivých častiach dňa. Tieto grafy je možné medzi jednotlivými dňami porovnávať a na základe priebehov hodnôt je možné nájsť pravidelnosti, ktoré sa v správaní daného užívateľa vyskytujú. Na základe činnosti užívateľa v minulosti je tak možné detekovať zmeny v správaní.

Na základe jednotlivých hodnôt v čase však nie je vhodné hľadať pravideľnosti chovania, pretože hodnoty vlastnosti v danom čase môžu medzi dňami výrazne kolísať. Ako príklad je možné uviesť situáciu, keď sa užívateľ jeden deň po príchode do práce pripojí na sieť, skontroluje svoju e-mailovú schránku a následne odpovedá na nevybavenú poštu. Nasledujúci deň však po príchode nemá žiadne žiadne správy, na ktoré je nutné odpovedať. Počas dňa však obdrží nové správy, ktorým zašle odpoveď. Množstvo odoslaných dát protokolom SMTP sa tak výrazne líši v jednotlivých časoch, v rámci celého dňa je však možné nájsť podobnosť. Hodnoty v jednotlivých časoch je preto potrebné transformovať

do podoby, aby nad nimi bolo možné vykonávať klasifikačné metódy. Túto transformáciu je možné vykonať rôznymi spôsobmi, napríklad aproximáciou pomocov polynómov alebo splajnov či vykonaním Fourierovej transformácie. Transformované dáta je možné následne spracovávať metódami dolovania znalostí z dát.

Ďalšou možnosťou, ako je možné analyzovať správenie používateľov v sieti je sledovanie jednotlivých akcii, ktoré užívateľ vykonáva. Akcia je definovaná ako trojica obsahujúca cieľ akcie, čas vykonania a počet prenesených bytov. Cieľom akcie sa rozumie kombinácia cieľovej IP adresy a cieľového portu. U každého používateľa je možné nájsť množinu cieľov, s ktorými pravidelne komunikuje a na základe počtu akcii užívateľa s daným cieľom a preneseného množstva dát je možné odlíšiť jednotlivých používateľov.

Kapitola 4

Návrh riešenia

Táto kapitola popisuje jednotlivé fázy riešenia problému analýzy správania užívateľov. Problém je riešený pomocou 2 prístupov. Prvý prístup využíva na analýzu štatistiku činnosti používateľov v sieti, pomocou druhého je popísané správanie na základe jednotlivých akcii vykonaných používateľom. Výstupy získané pomocou oboch prístupov sú následne analyzované nástrojom RapidMiner.

4.1 Spôsob uloženia dát

Pre potreby analýzy je potrebné získať dáta zo sieťovej prevádzky a spracovať ich do vhodnej podoby. Pre zjednodušenie boli dáta získavané z nešifrovanej siete, vďaka čomu je možné spracovať celú komunikáciu prebiehajúcu na sieti. Dáta sú zo siete zachytávané pomocou sondy Netflow [1]. Týmto spôsobom nie je potrebné spracovávať jednotlivé pakety, pretože sonda odosiela informácie o celých dátových tokoch prebiehajúcich na sieti. Pod dátovým tokom sa rozumie množina paketov, ktoré majú spoločné nasledujúce položky:

- zdrojová a cieľová IP adresa
- zdrojový a cieľový port
- vstupné rozhranie
- protokol na vrstve L3 (ICMP, IGMP, TCP, UDP)
- hodnota ToS

O jednotlivých tokoch sú zaznamenané informácie ako čas začiatku a ukončenia toku, dĺžka jeho trvania, počet odoslaných paketov a bytov. Sonda ukladá štatistiky o jednotlivých tokoch do súborov, ktoré je možné pomocou nástroja nfdump previesť do formátu csv, ktorý je jednoducho importovateľný do databázy.

4.2 Analýza na základe štatistík činnosti

Prvým spôsobom, akým je možné skúmať chovanie používateľov je rozbor štatistiky správania jednotlivých entít v sieti.

4.2.1 Spôsob uloženia dát

Tabuľka obsahujúca všetky informácie zachytené na sieti je príliš komplexná. Preto je potrebné dáta vhodne rozdeliť do viacerých tabuliek s ktorými je možné jednoducho pracovať.

Pre analýzu správania jednotlivých zariadení je ich potrebné najskôr identifikovať v sieťovej prevádzke. Ako identifikátor bola zvolená MAC adresa zariadenia. Táto voľba sa nejaví ako najvhodnejšia, pretože adresu MAC je veľmi jendoduché zmeniť. V takomto prípade, ak je útočník k sieti pripojený s podvrhnutou MAC adresou, by sa však mala prejaviť zmena v správaní zariadenia, čím bude odhalené, že útočník nie je užívateľ, za ktorého sa vydáva. Tabuľka `zariadenie` v ktorej sú uložené informácie o jednotlivých zariadeniach na sieti obsahuje:

- identifikátor zariadenia
- MAC adresu zariadenia
- počet dní, v ktorých zariadenie komunikovalo na sieti

Hodnota počtu dní je ukladaná, pretože na vytvorenie modelu správania je potrebné, aby bol k dispozícii dostatočný počet dní kedy zariadenie vykonávalo činnosť.

Jednotlivé analyzované vlastnosti správania, ktoré sú skúmané sú uložené v tabuľke `vlastnost`. Tabuľka obsahuje:

- identifikátor vlastnosti,
- skratku,
- popis vlastnosti,
- MySql dotaz, pomocou ktorého sú získané potrebné hodnoty z uložených tokov Netflow.

Hodnoty jednotlivých vlastností v čase sa ukladajú do tabuľky `hodnoty`. Tabuľka má nasledujúcu štruktúru:

- identifikátor
- id zariadenia
- id vlastnosti
- hodnota platná pre daný čas
- čas v dni
- dátum

Jednotlivé dni boli rozdelené na časové úseky 30 minút, pre ktoré sa hodnoty jednotlivých vlastností počítajú. Každá vlastnosť má preto denne 48 diskrétnych hodnôt.

4.2.2 Transformácia hodnôt

Ako je uvedené v kapitole 3, vykonávať analýzu nad samotnými hodnotami daných vlastností v čase nie je vhodné.

Jednou z možností, ako transformovať dáta je interpolácia hodnôt v čase polynómom [6]. Výstupom by bola postupnosť koeficientov polynómu, s ktorými by bolo možné ďalej pracovať. Túto možnosť som však zamietol, pretože s rastúcim počtom hodnôt rastie stupeň polynómu, ktorý interpoluje dané hodnoty, kvôli čomu by bolo nutné pracovať s veľkým množstvom koeficientov. Ďalšou možnosťou je interpolovať hodnoty pomocou splajnov. Táto možnosť bola odmietnutá z rovnakého dôvodu, keďže pri veľkom počte hodnôt je na interpoláciu potrebné veľké množstvo splajnov.

Hodnoty v čase je možné miesto interpolácie aproximovať pomocou jednoduchšieho polynómu, či inej funkcie. Keďže sa však priebehy hodnôt jednotlivých vlastností výrazne líšia, bolo by náročné určiť funkciu, pomocou ktorej sa majú hodnoty aproximovať.

Ďalšou možnosťou je využitie Diskrétnej Fourierovej transformácie[7]. Pomocou DFT je možné transformovať postupnosť hodnôt na postupnosť koeficientov konečnej kombinácie komplexných exponenciál zoradených podľa ich frekvencií. Jednotlivé hodnoty koeficientov je možné získať pomocou rovnice 4.1.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-2\pi i k n / N}, \qquad k = 0 \dots N-1 \tag{4.1}$$

Výstupné koeficienty sú komplexné čísla, preto treba porovnávať okrem ich reálnej hodnoty aj hodnotu imaginárnu. Vyššie koeficienty patria ku komplexným exponenciálam s vysokými frekvenciami, vo výsledku preto nemajú veľký vplyv a je možné ich zanedbať. Vďaka tejto vlastnosti som sa rozhodol pre účely analýzy zvoliť práve metódu Diskrétnej Fourierovej transformácie. V programe je použitá na výpočet koeficientov DFT knižnica FFTW [2].

4.3 Analýza akcii vykonávaných užívateľom

Ďalším spôsobom, akým je možné vykonávať rozbor správania používateľov v sieti je sledovanie jednotlivých akcii, ktoré sú vykonávané.

4.3.1 Spracovanie dát

Pre potrebu analýzy nie je potrebné zaznamenávať všetky akcie zachytené v sieťovej prevádzke. Analyzované sú iba služby, ktoré sú najčastejšie využívané, čím sa výrazne zníži množstvo skúmaných dát. Boli vybrané nasledovné služby:

- protokol HTTP a HTTPS (cieľový port 80 a 443)
- vzdialená komunikácia pomocou SSH (cieľový port 22)
- mailová komunikácia pomocou IMAP a SMTP (cieľový port 993 a 465)

Pomocou nástroja nfdump sú vyfiltrované toky, ktoré spĺňajú dané podmienky a následne sú exportované do formátu CSV, ktorý je jednoducho možné spracúvať strojovo.

Pomocou programu sú pre každé zariadenie spočítané výskyty akcii s jednotlivými cieľmi za deň a taktiež suma prenesených dát.

id	class	F_R_1_ibyte-cnt	F_R_2_ibyte-cnt	F_R_3_ibyte-cnt	F_R_4_ibyte-cnt
1	1	103033456	-69916808.111496	999719.283803	50520061.060591
2	1	9891819	-6629604.01729	2913148.645964	316087.600369
3	1	52854249	-32578328.895427	7557395.86749	-14118999.732535
4	1	4332746608	-3840449055.02298	2481372856.23806	-574285776.316454
5	1	787072	-80078.38397	-69468.660363	124509.857649
6	1	784398	75995.534024	-162886.456734	-193740.147151
7	1	4663949575	-3197434677.45204	-96949809.763309	3227139796.08107
8	1	869255106	-381988074.177586	280842136.268232	-112039367.358776
9	1	687066992	-95435597.857642	39353951.255393	-17877964.921686
10	1	129928678	101650319.538724	67043979.508043	12322842.081139
11	1	63290	43478.922403	48367.927076	25247.643497

Obrázek 4.1: Ukážka formátu výstupného súboru

4.3.2 Preklad IP adries na doménové mená

Väčšie domény môžu mať viacero serverov s rôznymi IP adresami. Pri identifikácii cieľa pomocou IP adresy by mohla nastať situácia, že viaceré pripojenia k jednej doméne by boli identifikované ako akcie s rôznymi cieľmi. Preto sú adresy IP pred identifikáciou cieľa pomocou DNS preložené na doménové mená. Taktiež sú zanedbané subdomény, cieľ akcie teda tvorí dvojica doména druhej úrovne a cieľový port. Cieľ, ktorého IP adresu nie je možné preložiť na doménové meno je identifikovaný adresou.

4.4 Export dát

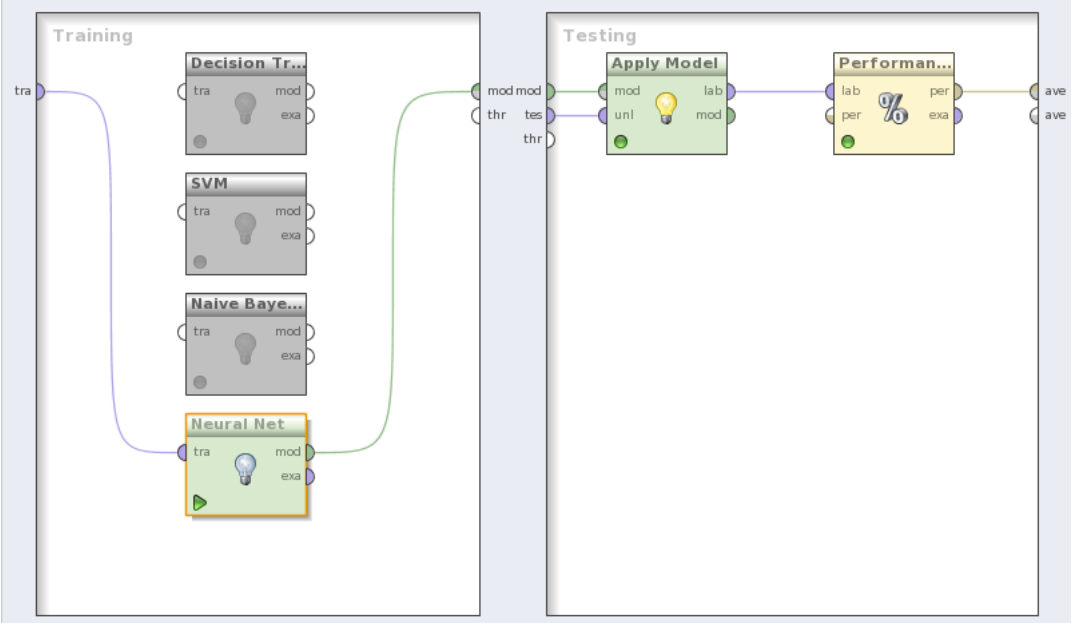
Výstupom programu je súbor formátu CSV, ktorý je možné použiť v dolovacom nástroji RapidMiner. Prvý riadok súboru tvorí hlavičku, ktorá určuje akú informáciu obsahuje daný stĺpec. V prvom stĺpci sa nachádza identifikácia záznamu, v druhom stĺpci číslo zariadenia, ku ktorému sa dáta viažu. Nasledujúce stĺpce sú označené v tvare `F_[R|I]_[1-5]_skratka-vlastnosti` obsahujúce jednotlivé reálne/imaginárne hodnoty koeficientov DFT pre danú sledovanú vlastnosť správania. Dáta získané sledovaním jednotlvých akcii majú hlavičku vo formáte `cnt-číslo_cieľa` a `bytes-číslo_cieľa` reprezentujúce počet akcii a celkové množstvo dát prenesených medzi zariadením a cieľom. Na nasledujúcich riadkoch sa nachádzajú hodnoty prislúchajúce daným stĺpcom oddelené čiarkami. Formát súboru je znázornený na obrázku 4.1.

4.5 Analýza dát nástrojom RapidMiner

Vygenerovaný CSV súbor som analyzoval pomocou nástroja na dolovanie znalostí RapidMiner [3] verzie 5.3.015 . V nasledujúcich podkapitolách sú popísané jednotlivé časti analýzy pomocou tohto nástroja.

4.5.1 Import dát

Nástroj RapidMiner umožňuje import zo súboru formátu CSV pomocou bloku *Read CSV*. Po zvolení oddeľovača hodnôt v importovanom súbore je potrebné určiť, ktoré riadky obsahujú komentáre, názvy a samotné dáta. Následne je nutné označiť význam jednotlivých stĺpcov, ktoré môžu obsahovať identifikátor zázanmu, triedu ku ktorej patrí záznam a samotné atribúty záznamu. Ako posledný krok je nutné nastaviť typy atribútov.



Obrázek 4.2: Schéma bloku Validation

4.5.2 Proces analýzy dát

Základom analýzy dát je blok *Validation*, ktorý slúži na vyhodnotenie úspešnosti klasifikácie pomocou použitej metódy. Vstupom bloku sú importované dáta pomocou nástroja *Read CSV*. Blok má nasledujúce výstupy:

- `model` - výstupný model, ktorý je závislý na použitej klasifikačnej metóde
- `training example set` - dáta, ktoré boli použité na trénovanie modelu sú bez zmeny odoslané na tento výstup
- `averagable` - výsledný performance vektor, ktorý určuje kvalitu klasifikácie a matica obsahujúca informácie o klasifikácii vzoriek do jednotlivých tried.

Na vyhodnotenie kvality modelu používa blok krížovú validáciu. Vstupné dáta sú rozdelené na N skupín. Proces vykoná N validácii, pričom pri každej je použitá iná skupina dát na testovanie, ostatné skupiny sú použité na trénovanie modelu. Schéma bloku sa nachádza na obrázku 4.2. V ľavej časti sa nachádzajú modely, ktoré je možné trénovať, z ktorých je aktívny iba jeden. V pravej časti prebieha testovanie, na vstup ktorého je privedený výstup modelu a testovacie dáta. Blok *Apply model* klasifikuje testovacie dáta pomocou natrénovaného modelu a blok *Performance* túto klasifikáciu vyhodnocuje a posiela výsledky na výstup.

Na trénovanie boli použité nasledujúce modely:

- Rozhodovací strom
- SVM
- Neurónová sieť
- Naivná Bayessovská metóda

Kapitola 5

Dosiahnuté výsledky a možnosti ďalšieho rozšírenia

Ako testovacia množina dát boli použité dáta zachytené Netflow sondou vo firme AEC, spol. s r.o.. Firma sa zaoberá poskytovaním software a služieb pre bezpečnosť dát a antivírovú ochranu, zamestnáva niekoľko desiatok zamestnancov. Dáta boli zbierané po dobu 2 mesiacov na jednej vlan obsahujúcej bezdrôtovú časť siete a tiež časť siete s pevným prenosovým médiom.

Najlepšie výsledky boli dosiahnuté použitím neurónovej siete s počtom tréningových cyklov 1000 a tempom učenia sa 0,3. V prípadoch, keď testovacie dáta pozostávali z hodnôt reprezentujúcich správanie jedného zariadenia a hodnôt, ktoré reprezentovali iné zariadenia, bola úspešnosť rozlíšenia daného zariadenia vždy minimálne 95%. Jeden z dosiahnutých výsledkov je možné vidieť v tabuľke 5.1. Je možné vidieť, že k správnej detekcii, že nejde o daného užívateľa došlo v 100% prípadov.

V ďalších testoch sa vo vstupnom súbore nachádzali dáta predstavujúce činnosť viacerých užívateľov a cieľom bolo klasifikovať jednotlivých užívateľov na základe ich správania. Najlepšie výsledky boli dosiahnuté taktiež pomocou neurónovej sieti s rovnakými parametrami. Príklad výsledku je možné vidieť v tabuľke 5.2. Celková úspešnosť klasifikácie je 94.05% +/- 2.02%.

Dobré výsledky boli dosiahnuté taktiež použitím rozhodovacieho stromu. Pre rovnakú sadu vstupných dát je možné vidieť výsledok v tabuľke 5.3. Celková úspešnosť klasifikácie bola 95.14% +/- 3.15%. Grafická reprezentácia rozhodovacieho stromu je na obrázku 5.1.

5.1 Možnosti budúcich rozšírení

Veľká nevýhoda implementovanej metódy klasifikácie užívateľov spočíva v tom, že je potrebné zozbierať dáta z celého dňa aby ich bolo možné analyzovať. Kvôli tomu je možné potencilneho záškodníka na sieti odhaliť až v čase, keď už svoju činnosť ukončil. Ako rozšíre-

	true True	true False	presnosť tried
pred. True	19	0	100.00%
pred. False	1	21	95.45%
odozva modelu	95.00%	100.00%	

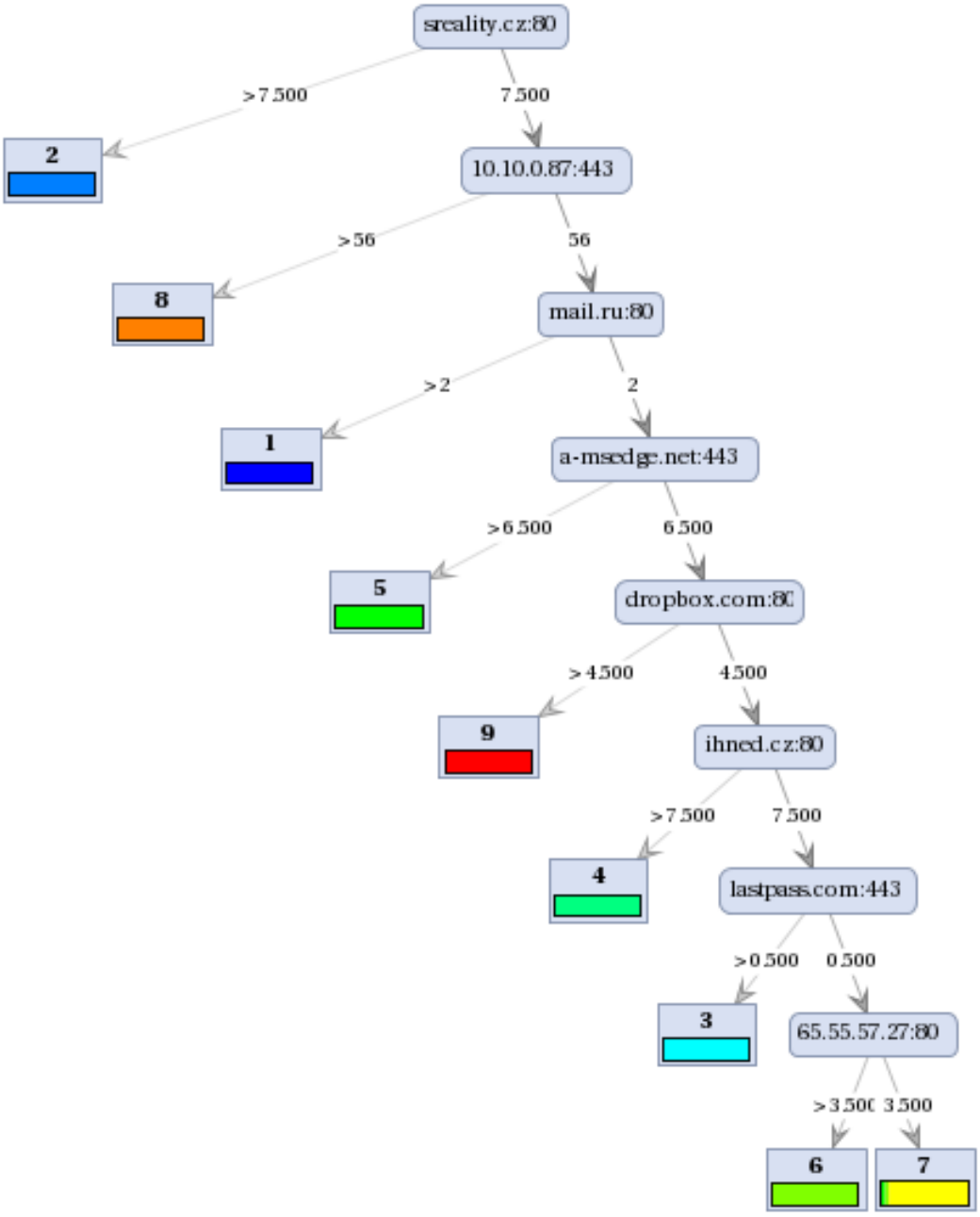
Tabulka 5.1: Výsledok odlíšenia jedného zariadenia

	true 1	true 2	true 3	true	true 5	true 6	true 7	true 8	true 9	presnosť
pred. 1	18	0	0	0	0	0	0	1	0	94.74%
pred. 2	0	19	0	0	0	0	0	0	0	100.00%
pred. 3	0	0	20	0	0	0	0	0	0	100.00%
pred. 4	0	0	0	20	0	1	1	0	0	90.91%
pred. 5	0	0	0	0	21	0	1	0	0	95.45%
pred. 6	1	0	0	0	0	20	1	0	0	90.91%
pred. 7	0	0	0	0	0	0	18	0	0	100.00%
pred. 8	0	1	0	0	0	0	0	19	1	90.48%
pred. 9	1	1	0	0	0	0	0	1	19	86.36%
odozva	90%	90.48%	100%	100%	100%	95.24%	85.71%	90.48%	95%	

Tabulka 5.2: Výsledok klasifikácie pomocou neurónovej siete

	true 1	true 2	true 3	true	true 5	true 6	true 7	true 8	true 9	presnosť
pred. 1	19	1	0	0	0	0	0	0	0	95.00%
pred. 2	0	20	0	0	0	0	0	0	0	100.00%
pred. 3	0	0	20	0	0	0	0	0	0	100.00%
pred. 4	0	0	0	20	0	1	0	0	0	95.24%
pred. 5	0	0	0	0	19	0	0	0	0	100.00%
pred. 6	1	0	0	0	0	19	2	0	0	86.36%
pred. 7	0	0	0	0	1	1	19	1	0	86.36%
pred. 8	0	0	0	0	0	0	0	20	0	100.00%
pred. 9	0	0	0	0	1	0	0	0	20	95.24%
odozva	95%	95.24%	100%	100%	90.48%	90.48%	90.48%	95.24%	100%	

Tabulka 5.3: Výsledok klasifikácie pomocou rozhodovacieho stromu



Obrázek 5.1: Nájdený rozhodovací strom

nie do budúcnosti by bolo možné implementovať klasifikátor tak, aby bolo možné analyzovať správanie užívateľa kedykoľvek počas dňa na základe jeho vývoja.

Prácu by bolo možné taktiež rozšíriť o modul, ktorý sleduje mobilitu používateľa v bezdrôtovej sieti na základe zmien prístupových bodov v čase, prípadne určením polohy na základe triangulácie síl signálov z jednotlivých prístupových bodov.

Kapitola 6

Záver

Cieľom práce bolo navrhnúť spôsob, akým je možné odlíšiť používateľov na sieti na základe ich správania. Výsledný klasifikátor je schopný užívateľov odlíšiť na základe charakteristických vlastností v ich komunikácii na sieti. Výsledok práce by pravdepodobne nebolo možné nasadiť v praxi, pretože úspešnosť klasifikácie je v niektorých prípadoch príliš nízka. Po implementácii navrhnutých rozšírení a zvýšení presnosti kvalifikátoru by však mohla mať práca praktické využitie.

Literatura

[1] Cisco IOS NetFlow.
http://www.cisco.com/c/en/us/products/ios-nx-os-software/ios-netflow/index.html.

[2] FFTW. http://www.fftw.org/.

[3] RapidMiner. http://rapidminer.com/products/rapidminer-studio/.

[4] *IEEE standard for information technology telecommunications and information exchange between systems-local and metropolitan area networks-specific requirements.* Institute of Electrical and Electronics Engineers, 2005.

[5] Benton, K.: The Evolution of 802.11 Wireless Security.
http://itffroc.org/pubs/benton_wireless.pdf , 2010.

[6] Fajmon, I., B.; Růžičková: Matematika3.
http://www.umat.feec.vutbr.cz/ novakm/matematika3.pdf,.

[7] Winigrad, S.: On computing the discrete Fourier transform.
http://www.ams.org/journals/mcom/1978-32-141/
S0025-5718-1978-0468306-4/S0025-5718-1978-0468306-4.pdf, 1976.